CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea

Rotem Sorek, Victor Kunin and Philip Hugenholtz

Abstract | Arrays of clustered, regularly interspaced short palindromic repeats (CRISPRs) are widespread in the genomes of many bacteria and almost all archaea. These arrays are composed of direct repeats that are separated by similarly sized non-repetitive spacers. CRISPR arrays, together with a group of associated proteins, confer resistance to phages, possibly by an RNA-interference-like mechanism. This Progress discusses the structure and function of this newly recognized antiviral mechanism.

Phages are the most abundant forms of life on Earth[1]. In seawater, an environment in which phage abundance has been extensively studied, it has been estimated that there are 5–10 phages for every bacterial cell[2]. Despite being outnumbered by phages, bacteria proliferate and avoid extinction by using various innate phage-resistance mechanisms, such as restriction enzymes and abortive infection[3]. In this Progress article, we describe the clustered, regularly interspaced short palindromic repeat (CRISPR) system, a recently discovered defence mechanism that is remarkable because it confers acquired phage resistance in bacteria and archaea. A hallmark of this system is the arrays of short direct repeats that are interspersed by non-repetitive spacer sequences.Additional components of the system include CRISPR-associated (CAS) genes and a leader sequence (FIG. 1a).

Brief history of CRISPR research

The first description of a CRISPR array was made in 1987 by Ishino and colleagues[4], who found 14 repeats of 29 base pairs (bp) that were interspersed by 32–33 bp non-repeating spacer sequences[5] and were adjacent to the isozyme-converting alkaline phosphatase (*iap*) gene in *Escherichia coli*. In subsequent years, similar CRISPR arrays were found in *Mycobacterium tuberculosis*[6], *Haloferax mediterranei*[7], *Methanocaldococcus jannaschii*[8], *Thermotoga maritima*[9] and other bacteria and archaea. The accumulation of sequenced microbial genomes allowed genome-wide computational searches for CRISPRs (the first such analysis was carried out by Mojica and colleagues[10] in 2000), and the most recent computational analyses revealed that CRISPRs are found in approximately 40% and 90% of sequenced bacterial and archaeal genomes, respectively[11,12] (BOX 1; TABLE 1).

In parallel with this initial analysis of the abundance of CRISPRs[13], Jansen and co-workers[14] identified four CAS genes that were almost always found adjacent to the repeat arrays. Subsequent studies initiated by Koonin and colleagues[15,16] and Haft and colleagues[17] detected 25–45 additional CAS genes in close proximity to the arrays. The same set of genes is absent from genomes that lack CRISPRs.

Several hypotheses for the function of CRISPRs have been proposed. In 1995, Mojica and co-workers[7] suggested that the repeats are involved in replicon partitioning, based on their observations that an increase in the copy number of the repeats in *Haloferax volcanii* results in altered replicon segregation. This effect, however, was not reproduced in similar experiments that were carried out in *M. tuberculosis*[14]. Based on the presence of several CRISPR loci in some genomes, Jansen and colleagues[14] suggested that CRISPRs are mobile elements, whereas Makarova and colleagues[15] suggested that the CRISPR system is involved in DNA repair, as many CRISPR-associated genes contained DNA-manipulating domains. In 2005, three research groups reported that the spacer sequences often contain plasmid- or phage-derived DNA, and proposed that CRISPRs mediate immunity against infection by extrachromosomal agents[18–20]. Bolotin and co-workers[20] also reported on a negative correlation between the sensitivity of bacteria to phage infection and the number of CRISPR spacers in their genome. Recently, Barrangou and colleagues[21–23] confirmed this hypothesis experimentally by showing that new spacers that were acquired following phage challenge confer resistance against the phage. Their discovery is discussed in more detail below.

Structural features of CRISPR systems

CRISPR arrays and CAS genes (which together form the CRISPR system) vary greatly among microbial species. The direct repeat sequences frequently diverge between species[14,24], and extreme sequence divergence is also observed in the CAS genes[16]. The size of the repeat can vary between 24 and 47 bp, with spacer sizes of 26–72 bp[12]. The number of repeats per array can vary from 2 to the current record holder, *Verminephrobacter eiseniae*[12], which has 249 repeats per array and, although many genomes contain a single CRISPR locus, *M. jannaschii* has 18 loci[8]. Finally, although in some CRISPR systems only 6, or fewer, CAS genes have been identified, others involve more than 20 (REF. 17). As discussed below, despite this diversity, most CRISPR systems have some conserved characteristics (FIG. 1a).

Repeats. In a single array, repeats are almost always identical with respect to size and sequence[14]. Despite being divergent between species, repeats can be clustered, based on sequence similarity, into at least 12 major groups[11]. Some of the larger groups contain a short (5–7 bp) palindrome — hence the word 'palindromic' in the CRISPR acronym[14]. These palindromes have been inferred to contribute to an RNA stem-loop secondary structure of the repeat[11], an hypothesis that is supported both by the

© 2008 Nature Publishing Group

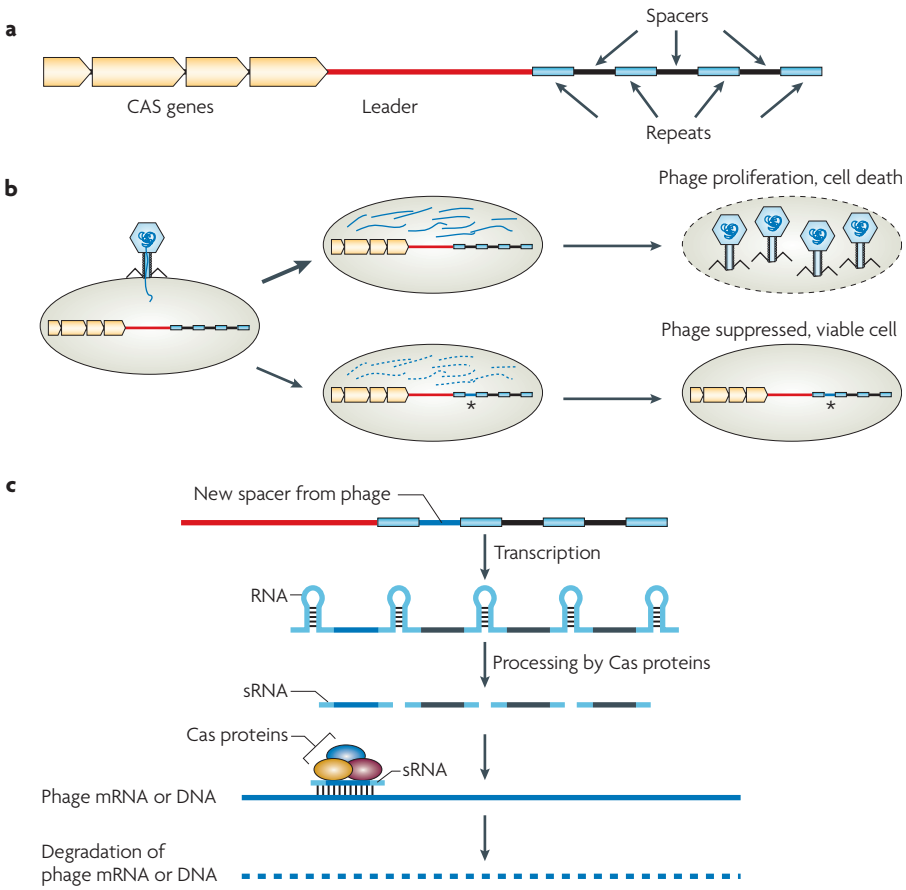

Figure 1 | **CRISPR structure and function. a** | Typical structure of a clustered, regularly interspaced short palindromic repeat (CRISPR) locus. **b** | CRISPRs acquire phage-derived spacers that provide immunity. Following an attack by a phage, phage nucleic acids proliferate in the cell and new particles are produced, leading to the death of the majority of the sensitive bacteria. A small number of bacteria acquire phage-derived spacers (marked by an asterisk), leading to survival, presumably by CRISPR-mediated degradation of phage mRNA or DNA. **c** | Putative, simplified model for CRISPR action. The repeat-spacer array is transcribed into a long RNA, and the repeats assume a secondary structure. Cas proteins recognize the sequence or structure of the repeats and process the RNA to produce small RNAs (sRNAs), each of which contains a spacer and two half repeats. The sRNAs, complexed with additional Cas proteins, base pair with phage nucleic acids, leading to their degradation. Putatively, this process is mediated by one or more of the Cas proteins. CAS, CRISPR-associated.

existence of compensatory mutations in the repeats that maintain the stem structure and by observations that the repeat-spacer array is transcribed into RNA[11,25–27]. For other repeat groups, evidence for RNA secondary structures is lacking. Apart from the structural feature, many repeats have a conserved 3′ terminus of GAAA(C/G). Both the structural features and the conserved 3′ motif have been suggested to act as binding sites for one or more of the CRISPR-associated proteins[11].

Spacers. In any CRISPR system, spacers are generally unique, with a few exceptions that are thought to have resulted from segmental duplications[12]. Similarity searches of various CRISPRs consistently showed that many spacers frequently match, with high sequence identity, to phages and other extrachromosomal elements[16,18–20,27]. Mojica and co-workers[18] studied 4,500 spacers from 67 microbial strains; 88 (2%) were similar to known sequences, and of these, more than 60% were similar to a sequence that is found within a known phage or plasmid. Comparable numbers were reported in a separate study in which 2,156 spacers were examined[20]. The observation that only 2% of all spacers match any known sequence presumably reflects the general under-sampling of phage-sequence space, and is in agreement with recent estimates of huge untapped phage environmental diversity[28]. Indeed, in lactic acid bacteria, such as *Streptococcus thermophilus*, for which more than a dozen phage genomes have been isolated and sequenced, approximately 40% of the spacers had a homologue matching either phage (75%) or plasmid (20%) sequences[20].

Spacers seem to be evenly distributed across the phage genomes and are derived both from the sense (coding) and antisense (non-coding) orientations[18,19,21,27], although one report suggested that there is a preference for spacers to be derived from one strand of the phage[20]. Two recent studies have reported that a short motif is present in phage genomes 1–2 nucleotides downstream of spacer-matching sequences[22,23]. This motif was proposed to be important for the recognition, or cleavage, of phage sequences by the CRISPR system. The recognition motif can vary between CRISPR systems, being AGAA and GGNG for the spacers found in the CRISPR1 and CRISPR3 loci of *S. thermophilus*, respectively.

Leader. A sequence of up to 550 bp is located 5′ to most CRISPR loci, directly adjoining the first repeat[14,27]. This common sequence has been denoted the 'leader' and is usually AT-rich[14]. Similar to repeats, leaders lack an open reading frame and are generally not conserved between species; however, if several CRISPR loci are found in the same chromosome their leaders can be conserved[8,29,30]. A new repeat-spacer unit is almost always added to the CRISPR array between the leader and the previous unit, which suggests that the leader could function as a recognition sequence for the addition of new spacers[19,21]. The leader has also been suggested to act as the promoter of the transcribed CRISPR array, as it is found directly upstream of the first repeat[25,26].

CAS genes. Two recent studies have characterized the large set of gene families that is associated with CRISPR arrays[16,17], and therefore, in this Review, only the general features of these genes are discussed. CRISPR systems have been divided into 7 or 8 subtypes; each subtype contains 2–6 different subtype-specific CAS genes. In addition, six core CAS genes (*cas1–6*) are associated with multiple subtypes, although the identity of *cas5* and *cas6* has not been agreed upon[16,17]. The *cas1* gene (NCBI COGs database code: COG1518) is especially noteworthy, as it serves as a universal marker of the CRISPR system (linked to all CRISPR systems except for that of *Pyrococcus abyssii*[16]). Additional genes that are more loosely associated with CRISPRs, such as members of the repeat associated mysterious protein (RAMP)[15,17]

© 2008 Nature Publishing Group

Box 1 | **Tools for CRISPR detection and analysis**

A growing interest in clustered, regularly interspaced short palindromic repeats (CRISPRs) has led to the development of different computer software and web resources for the analysis of CRISPR systems (TABLE 1). These tools include software for CRISPR detection, such as PILER-CR[49], CRISPR Recognition Tool[50] and CRISPRFinder[51]; online repositories of pre-analysed CRISPRs, such as CRISPRdb[12]; and tools for browsing CRISPRs in microbial genomes, such as Pygram[52]. The Institute for Genomic Research also provides a web page that displays the occurrence profile of all Cas proteins[17] for each available microbial genome. Among these tools, CRISPRdb is particularly notable as, apart from containing an automatically updated database of CRISPR arrays from published genomes (currently ~700 arrays in 232 genomes), it also provides various analysis tools that allow the extraction and alignment of specific repeats and spacers, as well as the flanking leader sequences. Despite this recent proliferation of tools for CRISPR analysis, there is still a need for tools that allow the combined analysis of CRISPR-associated (CAS) genes and CRISPRs, because most tools either focus on the repeat arrays or the related CAS genes. Reports that show the association between specific repeat types and specific CAS subsystems[11] highlight the need for such a combined web resource.

superfamily, occur only in genomes that contain CRISPR systems, but not necessarily near the CRISPR. Specific functional domains identified in Cas proteins include endonuclease and exonuclease domains, helicases, RNA- and DNA-binding domains, and domains that are involved in transcription regulation[14,16,17,31].

CRISPR is an antiphage defence system

Recently, Barrangou and co-workers[21] demonstrated experimentally that, in response to phage infection, bacteria integrate new spacers that are derived from phage genomic sequences, which results in CRISPR-mediated phage resistance (FIG. 1). These authors infected *S. thermophilus* with two different phages and recovered nine phage-resistant mutants. By sequencing the CRISPR1 locus, they showed that each of the phage-resistant mutants had independently acquired between one and four new repeat-spacer units at the leader-proximal end of the array, and that, in all cases, the spacers were derived from the genome of the challenging phage. If a spacer matched the phage sequence exactly (100% identity), the mutant was found to be phage resistant, but if one or more nucleotide changes were detected between the spacer and the phage sequence, bacteria were found to be phage-sensitive. Barrangou and colleagues[21] then inserted these resistance-conferring spacers into the CRISPR array of a phage-sensitive *S. thermophilus* strain, thereby causing it to become phage-resistant; finally, deletion of the acquired spacers caused the strain to become sensitive again.

Together, these results showed that inclusion of phage-derived spacers in CRISPR arrays confers resistance to phages. Interestingly, Barrangou and co-workers[21] noted that a small population of phages retained the ability to infect the resistant mutants. Further sequencing of the phage genomes revealed that the phages had mutated, so that their sequence was no longer identical to the spacers. Resistant phages that shared identical sequences with the spacers were also isolated, but the AGAA downstream recognition motif was mutated in their genome, which further strengthens the hypothesis that this motif is important for CRISPR function[22]. The selective pressure that is imposed by CRISPR on phages, therefore, leads to rapid changes in their genomes, and provides a glimpse into how CRISPR might be involved in driving the extremely high evolutionary rates that are observed in phages.

To begin to study the protein machinery that drives CRISPR function, Barrangou and colleagues[21] inactivated two subtype-specific CAS genes in a phage-resistant strain of *S. thermophilus*. The inactivation of *csn1* (REF. 17) (denoted *cas5* by Barrangou and colleagues[21]), which contains an endonuclease motif, resulted in loss of resistance, even in the presence of phage-derived spacers. Mutants that had a different *cas* gene inactivated (named *cas7* by Barrangou and colleagues[21]; might correspond to *cas2* or *csn2* according to the nomenclature of Haft and colleagues[17]) retained phage resistance if their CRISPR contained a phage-matching spacer, but were impaired in their ability to develop resistance to new phages, which might point to a role for this gene in acquiring new spacers[21].

A model for CRISPR activity

The exact mechanism by which CRISPR systems silence extrachromosomal DNA is not known, but a key observation was made by Tang and co-workers[25,26] who found, in

Table 1 | **Web resources for CRISPR analysis**

Resource and web page	Description	Refs
PILER-CR; http://www.drive5.com/pilercr/	A software tool for the detection of CRISPRs in microbial genomic sequences; based on local alignments in the genome that are represented by mathematical graphs*	49
CRISPR Recognition Tool; http://www.room220.com/crt/	A software tool for the detection of CRISPRs in microbial genomic sequences; based on the detection of exact k-mer matches that are separated by similar distances*	50
CRISPRFinder; http://crispr.u-psud.fr/crispr/	A software tool for the detection of CRISPRs in microbial genomic sequences; based on enhanced suffix arrays*	51
CRISPRdb; http://crispr.u-psud.fr/crispr/	Automatically updated database of CRISPR arrays in published microbial genomes; also contains CRISPR analysis tools that allow the alignment and comparison of repeats and spacers against the public databases	12
Pygram; http://www.irisa.fr/symbiose/projets/Modulome/article.php3?id_article=18	Visualization application that provides a graphical browser for studying repeats	52
TIGR Comprehensive Microbial Resource; http://rice.tigr.org/tigr-scripts/CMR2/genome_property.spl?subproperty=CRISPR%20region!&select_count=1	Provides a 'clickable' table that depicts, for each sequenced genome, the presence or absence of the 45 Cas protein families that are defined in Ref. 17	17

*This CRISPR (clustered, regularly interspaced short palindromic repeat) detection software applies post-processing filters to separate real CRISPR arrays from false predictions. BLAST, Basic Local Alignment Search Tool.

© 2008 Nature Publishing Group

species of *Archaeoglobus* and *Sulfolobus*, that the repeat-spacer array is transcribed into a single transcript, which is further processed into small RNA units, each of which is the size of a repeat plus a spacer. The cleavage position seems to reside in the middle of the repeat, which suggests that the processed small-RNA (sRNA) unit corresponds to a full spacer that is flanked by two half repeats (FIG. 1c). The existence of palindromic motifs within many repeats might indicate that the two half repeats attach to each other, with the spacer forming a loop.

The observation that CRISPRs are processed into sRNAs, as well as the assemblage of DNA- and RNA-manipulating protein domains within CAS genes, has led Makarova and colleagues[16] to suggest that CRISPR functions by an RNA-silencing (RNA interference (RNAi))-like mechanism. This mechanism has been well-characterized for its function as a defence against RNA viruses and transposable elements in eukaryotes[32]. In eukaryotic RNAi systems, long, double-stranded RNAs (dsRNA) of viruses are processed by a protein that is called Dicer into small interfering RNAs (siRNAs) that are 21–22 bp long. These siRNAs are converted into single strands by the RNA-induced silencing protein complex (RISC), and the RISC–siRNA complex identifies viral mRNAs by base pairing, leading to their degradation by another nuclease-denoted Slicer[33]. According to the RNAi hypothesis, the processed CRISPR spacers function as the microbial analogues of siRNAs. They bind to a RISC-like complex which comprises Cas proteins, and recognize the mRNA that is expressed from the foreign element by base-pairing, which results in subsequent degradation of the mRNA by other Cas proteins. Makarova and colleagues[16] further proposed that *cas3*, a protein that contains a helicase domain fused to an HD-nuclease domain, functions as the analogue of dicer and processes the transcribed repeat-spacer array into siRNAs. *cas4*, which encodes a RecB-like nuclease, was suggested to be the analogue of slicer[16]. A complication to this hypothesis stems from the observation that spacers can originate both from the sense and antisense strands of phage open reading frames[21]; a possible solution is that the spacers are first converted into dsRNA so that both strands participate in silencing[16]. Indeed, Lillestøl and colleagues[27] detected RNA transcripts that correspond to both strands of the CRISPR repeats in *Sulfolobus acidocaldarius*.

Evolution of CRISPR systems

CRISPR arrays can rapidly evolve, and CRISPR regions are often hypervariable between otherwise closely related strains[19]. A recent study revealed that in a nearly clonal population of a *Leptospirillum* species, which was identified by metagenomics in an acidophilic microbial biofilm, evolution of the spacer collection in CRISPR regions was fast enough to promote cell individuality[34]. As new spacers are almost always inserted at the 5′ end of the cluster next to the leader, the 'older' spacers (having greatest distance from the leader) are frequently common between isolates, whereas newer spacers are unique[19]. The deletion of repeat-spacer units is also frequently observed, which is necessary to prevent over-inflation of the CRISPR locus[12,19,22,23]; however, it is not clear whether such deletions occur actively or owing to passive homologous recombination. Rare duplications of repeat-spacer units were also observed[12].

On a higher evolutionary scale, CRISPR systems also greatly diversify. As indicated above, the repeats tend to vary between distantly related species, but exceptions are often noted. For example, the arrays in *E. coli* and *Mycobacterium avium* contain similar repeats, although these two organisms belong to different bacterial phyla[14]. This has been explained by horizontal gene transfer of CRISPR systems between organisms, a hypothesis that is supported by the phylogenetic trees of core CAS genes[16,17,24]. Gene transfer has been suggested to be mediated by megaplasmids, based on the identification of ten such plasmids that carry CRISPR arrays[24,35]. Interestingly, a CRISPR array was also found within a *Clostridium difficile* prophage, and it was suggested that the phage uses the CRISPR to limit the dispersal of competing phages[36].

Current and future applications

Strain typing. More than a decade before it was discovered that CRISPRs confer resistance to phages, Groenen and colleagues[37] had noticed that these loci are among the most rapidly evolving structures in the genome of *M. tuberculosis*, with strains varying in the number of repeats and the presence or absence of specific spacers. Based on this observation, Kamerbeek and colleagues[38] developed the spacer-oligotyping (also called spoligotyping) method for strain detection. In this method, probes for specific spacers are covalently bound to a membrane and hybridization patterns of labelled PCR products, which are primed from the CRISPR repeats, are measured (FIG. 2a). This has become the standard method for genotyping *M. tuberculosis* strains as part of ongoing efforts to control tuberculosis outbreaks[39,40], and is also used for the typing of *Corynebacterium diphtheriae*[41]. Non-spoligotyping-based methods for strain typing using CRISPR arrays have been used to study *Campylobacter jejuni*, *Thermotoga neapolitana* and other bacterial strains[42,43], and Russell and colleagues recently filed a patent application on CRISPR-based methods to type *Lactobacillus* spp. strains[44].

Engineered defence against viruses. Many industries that are reliant on bacteria, such as the dairy and wine industries, are concerned about phage infection. Owing to the high costs that are associated with phage-mediated culture losses, the dairy industry invests heavily in efforts to combat phage infection of dairy bacteria[3]. CRISPRs might offer a partial solution to this problem — by artificially adding spacers that are derived from conserved regions of known phages to the CRISPR array of industrial bacteria, manufacturers could boost the immunity of their starter cultures against known phages (FIG. 2b). A recent patent application based on this concept has been filed by Horvath and colleagues[45].

Selective silencing of endogenous genes. As noted above, it has been proposed that the CRISPR system is analogous to the eukaryotic RNAi system and that the spacers function as prokaryotic siRNAs by base-pairing with foreign mRNAs and promoting their degradation[16]. Should this hypothesis be confirmed, then manipulated CRISPR systems could revolutionize microbial-physiology research, as they would allow selective gene knockdown without manipulation of the original microbial genome. Instead of knocking out the gene of interest, which is usually labour intensive, the same effect could be achieved by transforming a CRISPR-bearing plasmid into the organism of choice, with one of the spacers being changed to match the studied gene (FIG. 2c). Moreover, the array nature of CRISPRs could allow the simultaneous knockdown of multiple endogenous genes. Similar RNAi-based applications have revolutionized eukaryotic genetic studies; we envisage that CRISPRs would have a similar impact in the field of microbial genetics.

Outlook

Despite the recent advances in understanding the role of CRISPRs in microbial genomes, the mechanisms that underlie CRISPR function are uncharacterized and the current

© 2008 Nature Publishing Group

hypotheses mainly rely on educated guesses that are based on bioinformatic analyses. Fundamental questions, such as how new spacers are selected and inserted, how silencing of foreign DNA and RNA is achieved and whether different CRISPR systems contain different functionalities, are all expected to be addressed in the near future by the growing number of groups who are studying this system. Other questions that could be addressed in the future, following extensive research on the system, are discussed below.

The widespread occurrence of CRISPR systems in nearly half of all sequenced bacterial genomes points to their efficiency in providing protection against phage attacks. However, phages are the most abundant biological entities on Earth[1,46], and so it is plausible that phages have evolved various mechanisms to escape or inhibit CRISPRs. In fact, the high rates of evolution that are observed in CRISPR repeats and their associated proteins indicate that an 'arms race' between phages and CRISPR systems might be occurring, in which mutations in the CRISPR systems mediate escape from CRISPR shut-down mechanisms that are encoded by phages. If this hypothesis is correct, we would expect reports of phage-encoded anti-CRISPR systems. Hints that such a system exists can be found in the report by Peng and co-workers[47], in which they describe a *Sulfolobus* protein that specifically binds to the CRISPR DNA and induces an opening of the structure near the centre of the repeat. We performed a homology search of this protein against all available microbial genomes, and found that its homologues are mainly found in bacterial prophages (Sorek R., unpublished observations). We therefore propose that this protein might constitute part of an anti-CRISPR system that is encoded by phages; its exact role in this system is unknown.

The proposed analogy between the CRISPR system and eukaryotic RNAi raises another possible important role for CRISPRs. In eukaryotes, RNAi functions both in silencing foreign elements through siRNAs, as well as endogenous gene regulation through genome-encoded micro-RNAs. Analogously, it is possible that CRISPR systems regulate endogenous functions in different bacteria. Indeed, 7–35% of the spacers found in CRISPR arrays have homologues in the chromosomal DNA, which may indicate that CRISPR is being used to regulate the expression of chromosomally derived genes[18,20,23]. Moreover, the *devTRS* operon in *Myxococcus xanthus*, which encodes genes that are essential for spore differentiation

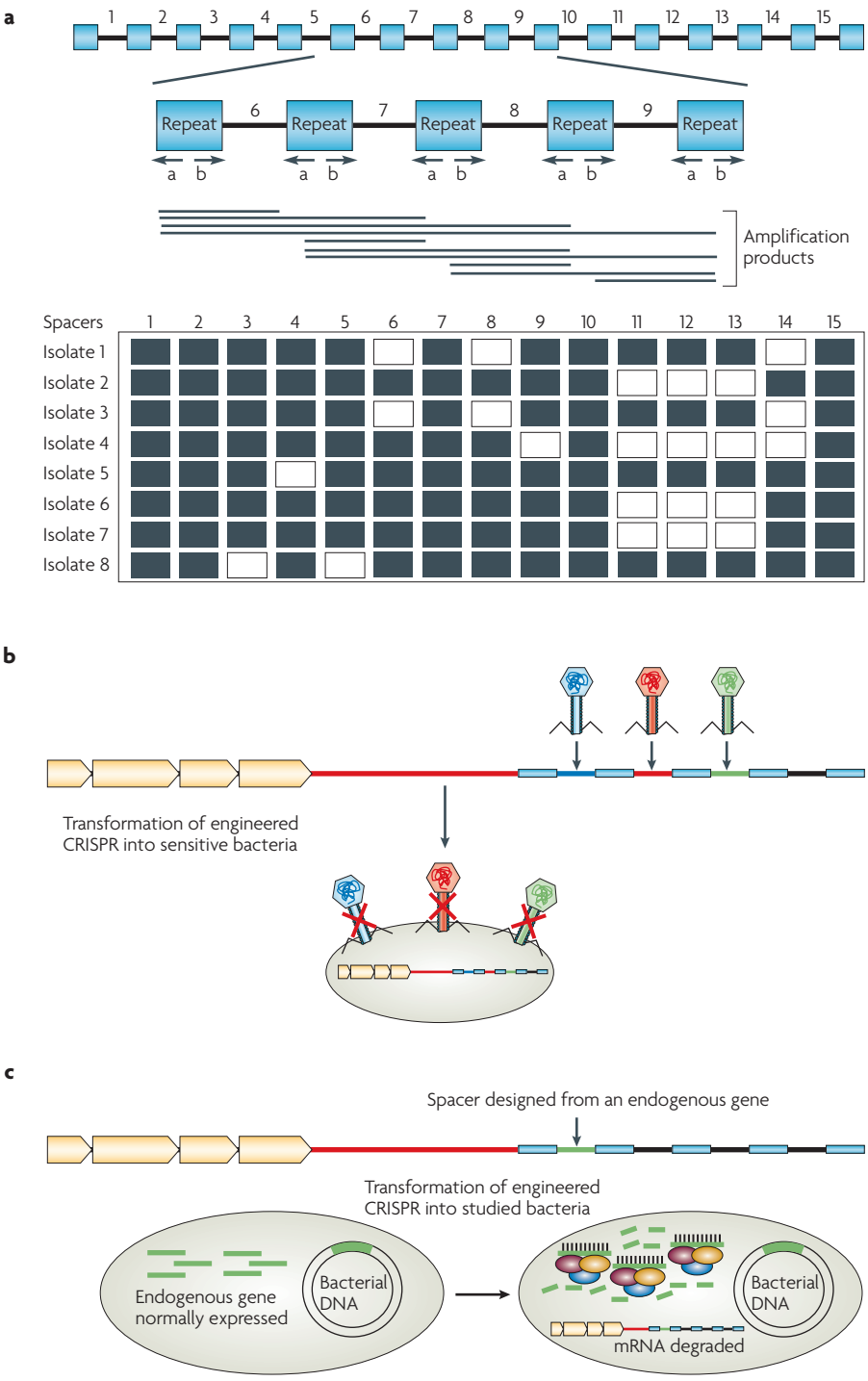

Figure 2 | **Applications of CRISPRs. a** | Spoligotyping. Labelled primers (a and b) are designed from the repeat region to amplify the clustered, regularly interspaced short palindromic repeat (CRISPR) array. Probes that match known spacers are printed on a membrane, and the amplified products for each isolate are hybridized. Black boxes represent the presence of a spacer and white boxes represent the absence of a spacer. Isolates 1 and 3 belong to the same strain, and isolates 2, 6 and 7 belong to the same strain. **b** | Engineering of phage resistance into sensitive industrial bacteria. Sequences from known phages are inserted as spacers into a CRISPR array and the CRISPR system is then transformed into bacteria. **c** | Silencing of endogenous genes as an alternative to knockout methods. Fragments from a chromosome-encoded gene (green) are engineered into a CRISPR array as spacers. If, as suggested, the CRISPR system indeed functions by the silencing of RNA[16], this might lead to silencing of the endogenous gene. Part **a** modified, with permission, from REF. 38 © 1997 American Society for Microbiology.

© 2008 Nature Publishing Group

inside fruiting bodies, is co-transcribed within a CRISPR operon, with DevS being a *bona fide* Cas5 protein[17,48]. This might be an example of a CRISPR system that regulates an endogenous mechanism.

Conclusions

Previously considered to be a simple family of repetitive elements, the CRISPR system has begun to take centre stage in our understanding of acquired phage resistance in prokaryotes. The widespread presence of this system in many bacterial and archaeal phyla, as well as its extreme diversity, suggest that it may be one of the most ancient antiviral defence systems in the microbial world[16]. Future studies are expected to define how CRISPR functions and elucidate the role of this system in host–phage co-evolution.

Rotem Sorek, Victor Kunin and Philip Hugenholtz are at the Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA.
Correspondence to R.S.
e-mail: rsorek@lbl.gov

doi:10.1038/nrmicro1793
Published online 24 December 2007

1. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284 (2005).
2. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).
3. Sturino, J. M. & Klaenhammer, T. R. Engineered bacteriophage-defence systems in bioprocessing. *Nature Rev. Microbiol.* **4**, 395–404 (2006).
4. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* **169**, 5429–5433 (1987).
5. Nakata, A., Amemura, M. & Makino, K. Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *J. Bacteriol.* **171**, 3553–3556 (1989).
6. Hermans, P. W. *et al.* Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect. Immun.* **59**, 2695–2705 (1991).
7. Mojica, F. J., Ferrer, C., Juez, G. & Rodriguez-Valera, F. Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol. Microbiol.* **17**, 85–93 (1995).
8. Bult, C. J. *et al.* Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073 (1996).
9. Nelson, K. E. *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
10. Mojica, F. J., Diez-Villasenor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.* **36**, 244–246 (2000).
11. Kunin, V., Sorek, R. & Hugenholtz, P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.* **8**, R61 (2007).
12. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172 (2007).
13. Jansen, R., van Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of a novel family of sequence repeats among prokaryotes. *OMICS* **6**, 23–33 (2002).
14. Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
15. Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B. & Koonin, E. V. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.* **30**, 482–496 (2002).
16. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7 (2006).
17. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1**, e60 (2005).
18. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).
19. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
20. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).
21. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
22. Deveau, H. *et al.* Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.* 7 Dec 2007 (doi:10.1128/JB.01412-07).
23. Horvath, P. *et al.* Diversity, activity and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* 7 Dec 2007 (doi:10.1128/JB.01415-07).
24. Godde, J. S. & Bickerton, A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.* **62**, 718–729 (2006).
25. Tang, T. H. *et al.* Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. USA* **99**, 7536–7541 (2002).
26. Tang, T. H. *et al.* Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* **55**, 469–481 (2005).
27. Lillestøl, R. K., Redder, P., Garrett, R. A. & Brügger, K. A putative viral defence mechanism in archaeal cells. *Archaea* **2**, 59–72 (2006).
28. Edwards, R. A. & Rohwer, F. Viral metagenomics. *Nature Rev. Microbiol.* **3**, 504–510 (2005).
29. Klenk, H. P. *et al.* The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370 (1997).
30. Smith, D. R. *et al.* Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155 (1997).
31. Ebihara, A. *et al.* Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci.* **15**, 1494–1499 (2006).
32. Hannon, G. J. RNA interference. *Nature* **418**, 244–251 (2002).
33. Sontheimer, E. J. Assembly and function of RNA silencing complexes. *Nature Rev. Mol. Cell Biol.* **6**, 127–138 (2005).
34. Tyson, G. W. & Banfield, J. F. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.* 26 Sep 2007 (doi: 10.1111/j.1462-2920.2007.01444.x).
35. Greve, B., Jensen, S., Brügger, K., Zillig, W. & Garrett, R. A. Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. *Archaea* **1**, 231–239 (2004).
36. Sebaihia, M. *et al.* The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nature Genet.* **38**, 779–786 (2006).
37. Groenen, P. M., Bunschoten, A. E., van Soolingen, D. & van Embden, J. D. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol. Microbiol.* **10**, 1057–1065 (1993).
38. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914 (1997).
39. Crawford, J. T. Genotyping in contact investigations: a CDC perspective. *Int. J. Tuberc. Lung Dis.* **7**, S453–S457 (2003).
40. Brudey, K. *et al. Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.* **6**, 23 (2006).
41. Mokrousov, I., Limeschenko, E., Vyazovaya, A. & Narvskaya, O. *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. *Biotechnol. J.* **2**, 901–906 (2007).
42. Schouls, L. M. *et al.* Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J. Clin. Microbiol.* **41**, 15–26 (2003).
43. DeBoy, R. T., Mongodin, E. F., Emerson, J. B. & Nelson, K. E. Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *J. Bacteriol.* **188**, 2364–2374 (2006).
44. Russell, W. M., Barrangou, R. & Horvath, P. Detection and typing of bacterial strains. US Patent Application 20060199190 (2006).
45. Horvath, P., Barrangou, R., Fremaux, C., Boyaval, P. & Romero, D. International Patent Application 2007025097 (2007).
46. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361 (2005).
47. Peng, X. *et al.* Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. *J. Bacteriol.* **185**, 2410–2417 (2003).
48. Viswanathan, P., Murphy, K., Julien, B., Garza, A. G. & Kroos, L. Regulation of *dev*, an operon that includes genes essential for *Myxococcus xanthus* development and CRISPR-associated genes and repeats. *J. Bacteriol.* **189**, 3738–3750 (2007).
49. Edgar, R. C. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, 18 (2007).
50. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
51. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52–W57 (2007).
52. Durand, P., Mahé, F., Valin, A. S. & Nicolas, J. Browsing repeats in genomes: Pygram and an application to non-coding region analysis. *BMC Bioinformatics* **7**, 477 (2006).

Acknowledgements

The authors thank H. Garcia Martin, M. J. Blow, A. Visel and G. Tyson for helpful discussions. This work was performed under the auspices of the US Department of Energy, Office of Science, Biological and Environmental Research Program at the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

DATABASES

Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Campylobacter jejuni | *Clostridium difficile* | *Corynebacterium diphtheriae* | *Escherichia coli* | *Haloferax volcanii* | *Methanocaldococcus jannaschii* | *Mycobacterium avium* | *Mycobacterium tuberculosis* | *Myxococcus xanthus* | *Pyrococcus abyssii* | *Streptococcus thermophilus* | *Sulfolobus acidocaldarius* | *Thermotoga maritima* | *Thermotoga neapolitana* | *Verminephrobacter eiseniae*
NCBI COGs database: http://www.ncbi.nlm.nih.gov/COG/cas1

FURTHER INFORMATION

Rotem Sorek's homepage: http://durian.jgi-psf.org/~rsorek/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2008 Nature Publishing Group